DATE: March 27, 1975

TO : Distribution

FROM : M. Pavkovic

SUBJECT: The problem of overlapping peaks from an advanced point of view

ABSTRACT

Recently, a new approach for the utilization of GC-MS data has been proposed. The main novelty consists of subjecting all of the eluting components from the gas chromatograph to the MS scans. The obtained huge volume of new information enables a chemist to get a much more detailed look in the chemical composition of a sample than previously. Unfortunately, there is no evidence that the new technique will make the problem of overlapping peaks obsolete, although it will certainly take away some of its edge.

This is perhaps a good time for a brief critical review of the field. An advanced viewpoint is adopted, because the "advanced" viewpoints tend to be unifying, and it is in this form, the author believes, the material should be handed over to the chemists and to the programmers.

INTRODUCTION

Recently, a novel approach for the utilization of GC-MS data has been proposed [1, 2, 3].

In this approach the mass spectra of all eluting components from the gas chromatograph, even very minor ones, are collected in a single experimental run. This technique places a wealth of new information into the hands of a chemist, who is now capable, with the help of a computer, to get a much closer and detailed look into the true chemical composition of a sample. Unfortunately, the new technique does not obviate the necessity of dealing with the problem of overlapping peaks, although it does hold the promise that this problem will now appear with much less severity than previously. An example of a situation where the problem of overlapping peaks can hardly be avoided is provided by two eluting components with similar chromatographic retention indices which are characterized by only one prominent ion in their respective mass spectra, and which ion happens to be the same in both spectra. Then, the only available ion that can betray the presence of these two components is the one in common, and if the retention indices lie sufficiently close, the resulting peak in the (retention index, ion current) plot will be a composite one, asking for a suitable resolution technique to cope with the situation.

There exists a number of resolution techniques currently in use. They can be devided into two more or less distinct classes: those with a well-defined mathematical and conceptual basis and those of more heuristic nature.

CLASS A

This class covers those techniques which have a sound mathematical and conceptual basis. The class A is characterized by the ability to cope, at least in principle, with all possible situations that can arise in practice. In other words, any peak shape no matter how complicated, can be subjected to an analysis whose ultimate result must be a unique sequence of elementary peaks which approximate, in a well-defined mathematical sense, the original composite peak.

The basic underlying assumption in the class A is that the "elementary" peaks can be represented by a family of analytic functions such as the gaussian functions, shew-gaussian functions, Poisson Distribution functions, and others. We will have

$$f(x) = a_1 \varphi_1(x,\alpha) + a_2 \varphi_2(x,\alpha) + \ldots + a_n \varphi_n(x,\alpha),$$

where $\varphi_i(x,\alpha)$ are the elements of this family, a_K the coefficients to be determined by the fit and $f(x)$ represents the shape of the composite peak. α stands for a fixed set of free parameters whose values must also be determined uniquely from the fit. Approximating $f(x)$ in terms of $\varphi_i(x,\alpha)$ functions means making the difference

$$(1) \qquad h(x) = f(x) - a_1 \varphi_1(x,\alpha) - \ldots - a_n \varphi_n(x,\alpha)$$

as small as possible. Mathematics knowns at least two ways of giving the vague notion of smallness a rigorous meaning. One is based on the concept of norm, while the other utilizes a more general notion of a functional. Both notions associate a number, or a set of numbers, to a function, but they do this in a somewhat different manner.

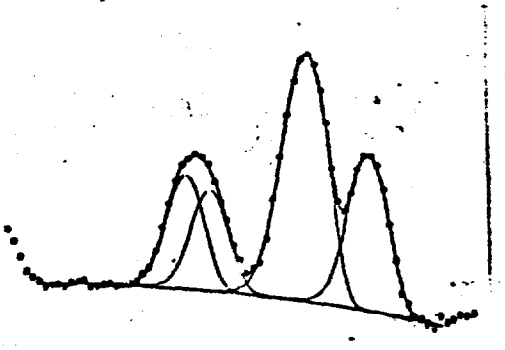
Least-Squares method: A common definition of a norm on a space of functions is the integral

$$(2) \qquad \|h\| = \left[\int |h(x)|^2 dx \right]^{1/2}$$

We say that the function h is small if the norm $\| h \|$, defined by (2), is small. In this context the approximation (or resolution, as chemists call it) of a composite peak $f(x)$ in terms of a fixed set of elementary peaks $\varphi_i(x,\alpha)$ means determining the parameters a_i and α in such a way to make the integral

$$(3) \qquad \int \left| f(x) - a_1\varphi_1(x,\alpha) - a_2\varphi_2(x,\alpha) - \dots - a_n\varphi_n(x,\alpha) \right|^2 dx$$

as small as possible. This is the famous least-squares approximation whose graphical illustration is given below.



A beautiful example of the least-squares approximation. The data are resolved into the four gaussian functions plus a polynomial background.

From the viewpoint of numerical mathematics the least squares approximation is particularly suitable approach to a fitting problem if the functions $\varphi_i(x,\alpha)$ form so called orthogonal set. The orthogonality is defined by the requirement that the integrals

$$\int \varphi_i(x,\alpha)\varphi_j(x,\alpha)\,dx$$

all vanish for $i \neq j$. Unfortunately, the functions that can conceivably be used in the problem of resolution of composite peaks are not orthogonal. This can be easily verified on the example of gaussian peaks where the integrals

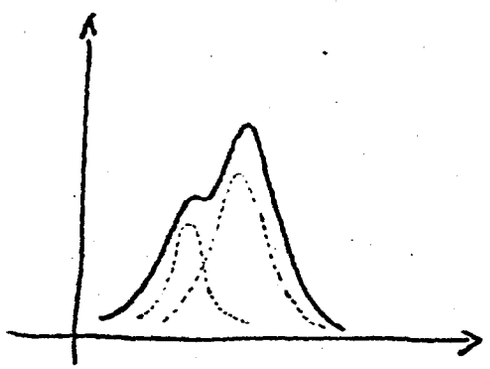
$$\int \left[c_i e^{-b_i (x-x_i)^2}\right]\left[c_j e^{-b_j (x-x_j)^2}\right] dx \quad , \; b_i, b_j, c_i, c_j > 0$$

refuse to vanish for any finite choice of the peak parameters. Physically, it is quite obvious why the peaks $\varphi(x,\alpha)$ cannot form an orthogonal set. They are all positive functions (number of ions variable cannot possibly accept a negative value!), and an integral of a product of positive functions is always positive itself, never zero or negative.

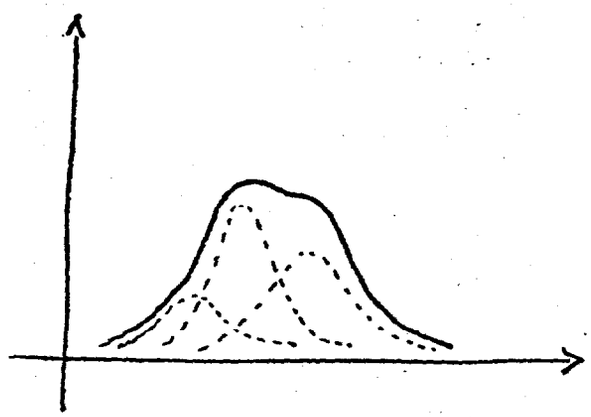
Without having the benefit of dealing with an orthogonal family of functions, the numerical mathematics usually faces great difficulties in pursuing the least squares approach. The trouble is that the integrals (3) usually display a very complicated functional dependence upon the parameters a_i and α, and it becomes a non-trivial problem to determine for which values of these parameters the integrals reach their minimum. It is also known that the outcomes of the iteration procedures which are employed in the least squares calculations are notoriously sensitive upon the correct choice of

the starting set of parameters. An unfortunate choice may lead to divergences or to an inordinately large number of iterations. Or, one may end up with a "false" solution, lacking reasonable physical interpretation.

In summary, the method of least squares is mathematically well-defined and conceptually easy to comprehend. These are the obvious advantages. When applied to the problem of overlapping peaks the least squares method has the unfortunate tendency of running into the difficulties of computational nature. Immediate dangers are divergences, slow convergence and misleading solutions. When contemplating the use of least squares in the problem of peak resolution, one should first try a judicious set of the family of functions $\varphi_i(x,\alpha)$, representing the elementary peaks. A desirable set is the one which makes the functional dependence of the integrals (3) upon the peak parameters reasonably simple. Neeldess to say, the functions $\varphi_i(x,\beta)$ must also resemble the peaks actually observed in the laboratory. Next, a suitable heuristic procedure should be devised, whose purposeshould be to make the initial guess of the peak parameters as close as possible to their true values. This may not be an easy task, if the original composite peak displays a tendency of hiding its internal structure.



A situation where it is not difficult to make a correct initial guess of the values of the elementary peak parameters



A situation where it is difficult to guess the approximate values of the elementary peak parameters.

Resolution of peaks in the sense of "weak" topology:

Norm represents one way of introducing the concept of "smallness" or "proximity" in a space of functions. The associated topology is called "strong" topology. There exists another way of introducing topology in the space of functions called "weak" topology. The two are related, but the relationship cannot be properly explained without invoking concepts of more advanced mathematics. This is not our intention here, in this brief review. Let us consider a set of functions $\tau_i(x)$, which can be finite or infinite, orthogonal or not. Call this set $\mathcal{T}$. To each function $h(x)$ we associate the set of numbers h_i, defined by

$$h_i = \int h(x)\,\tau_i(x)\,dx \quad .$$

Now, we can say that h(x) is small on the set $\mathcal{T}$, if the sequence of numbers h_i is small. If h(x) is the difference (1) between the composite peak function $f(x)$ and the linear combination of elementary peaks $\varphi_i(x,\alpha)$, the sequence of inequalities

$$h_i \leq \epsilon_i \quad (\epsilon_i \text{ small, say } < 10^{-3})$$

defines in a mathematically rigorous sense the degree of proximity in which the elementary peaks approximate the function $f(x)$. We can call this method the resolution of composite peaks in the sense of weak topology.

The choice of functions $\tau_i(x)$ is dictated by the mathematical convenience and physical considerations. For example, when dealing with the gaussian peaks, it is convenient to use polynomials for $\tau_i(x)$

An example:

Consider a composite peak $f(x)$, and let us restrict our efforts to the resolution of this peak into not more than two gaussian peaks. Furthermore, for the sake of simplicity, we may assume that these two peaks have the same width and the same height, and differ only in location. At the end, we are left with the three free parameters: the two peak locations plus the peaks height. For τ_i functions we use the powers

$$\tau_i = x^i, \qquad i = 0, 1, 2.$$

The calculations go as follows:

Normalization of Elementary Peaks:

$$N_j \int e^{-b_j (x-x_j)^2} dx = 1$$

Definition of the Moments:

$$I_K : \qquad I_K = \int f(x)\, x^K dx$$

$$\ldots \quad c_1 N_1 e \qquad + \quad c_2 N_2 e^{-2(x-x_2)}$$

$$c_1 = c_2 = c$$

$$b_1 = b_2 = b$$

$$x_1 \neq x_2$$

$$I_0 = c_1 + c_2 = 2c$$

$$I_1 = c_1 N_1 \int x\, e^{-b_1 (x-x_1)^2} dx \quad + \quad c_2 N_2 \int x\, e^{-b_2 (x-x_2)^2} dx$$

$$I_2 = c_1 N_1 \int x^2 e^{-b_1 (x-x_1)^2} dx \quad + \quad c_2 N_2 \int x^2 e^{-b_2 (x-x_2)^2} dx$$

A simple calculus leads to

$$I_0 = 2c$$

$$I_1 = c\,(x_1 + x_2)$$

$$I_2 = c\,(x_1^2 + d_1) + c\,(x_2^2 + d_2) \qquad , \; d_j = N_j^{-1} \frac{d}{db_j} N_j$$

or ,

$$c = I_0/2 \qquad x_{1,2} = I_1/I_0 \mp \sqrt{I_2/I_0 - I_1^2/I_0^2 - (d_1 + d_2)/2}$$

Without losing generality we can place the origin of the coordinate system in the center of the original composite peak, which is obviously a symmetrical peak (otherwise we would not be able to decompose it into the two gaussians of the same width and the same height!). This will set I_1 to zero. By acknowledging that

$$N_j^{-1} \frac{d}{db_j} = \frac{1}{2b_j}$$

and that $b_1 = b_2 = b$

we obtain finally

$$c = I_0/2$$

$$x_1 = -\sqrt{I_2/I_0 - 1/2b} \qquad x_2 = \sqrt{I_2/I_0 - 1/2b}$$

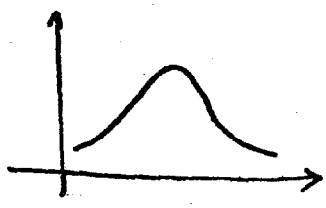
This example illustrates how the method works. In a more complicated situation we will have more free parameters, and will need correspondingly larger number of ξ_i functions. Instead of using simple powers of x we may also contemplate the use of more sophisticated ξ_i functions, perhaps a subset of an orthogonal set of functions.

The least squares approach and the approach via the weak topology all but exhaust the techniques of peak resolution that we classified under A. The advantages of starting with a well-defined mathematical procedure are obvious. It is also important that no peaks lie beyond the reach of these procedures. At least in principle, any reasonably behaved positive definite function can be expressed as a linear superposition of a sufficiently large number of gaussian functions, Poisson Distribution functions, or some other sufficiently rich class of functions.
After the process of peak resolution is completed, the obtained sequence of elementary peaks is amenable to a sensible chemical interpretation. Using various criteria, one can separate chemically significant peaks from the peaks originating from column bleed, electronic noise or some other undesirable source.

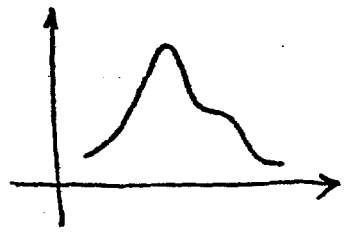
The disadvantages of the procedures from the class A consist mainly in the computational difficulties which are a growing function of the number of free parameters that have to be determined. The algorithms either show the tendency to fail on more difficult cases, or require large execution times. They may also lead to "spurious" or "false" solutions.

CLASS B

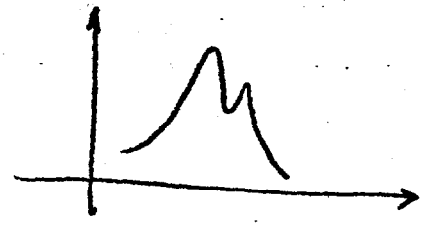
This class contains various techniques for peak resolution which are not based on solid mathematical reasoning. These are heuristic techniques. From the outset their scope is limited. For example, a particular heuristic algorithm may require that a composite peak displays a clear dip as a marker for the place where two peaks overlap. If the constituent peaks overlap too strongly, and no such dip is visible, the algorithm will not work. It was not intended to work with such a case, in the first place. And so, some potentially interesting peak shapes are left out from the beginning. This is a universal weakness of the heuristic models. The advantages of the algorithms from the class B lie in their simplicity. As a rule they perform well and produce fast results when applied to the cases which lie well within their domain of competence. Perhaps a combination of several heuristic methods would be desirable. To fix the ideas consider three different algorithms A1, A2 and A3, designed to deal with the following three classes of composite peak shapes.



SYMMETRIC SHAPE
Algorithm #1



SHOULDER SHAPE
Algorithm #2



DIP SHAPE
Algorithm #3

After a composite peak was scanned, a special procedure would determine in which class of shapes this peak belongs. Then, the ocrresponding algorithm would be conscribed to perform the peak resolution. In the general case the special procedure may involve the whole library of characteristic peak shapes. The search through this library may display some of the characteristics of the pattern matching algorithms. A brief outline of the overall algorithm may look something like this.

(1) First, use the least squares method to approximate the data with a polynomial of appropriately high degree. Evidently the number of available data points will place restrictions on the degree of the polynomial. Note that the use of least squares in the case of polynomials is a well-known and straightforward computational procedure.

(2) Determine the number and locations of the local maxima and minima, points of inflection and other critical parameters that may be used to determine the shape of a peak. Since we deal here with the polynomials, this information comes from a large and well-known chapter of classical analysis. The characteristics of peak shapes will be reflected in the restrictions imposed on the coefficients of the approximating polynomials. This is a convenient way of coding the peak shapes, namely converting purely geometrical entities into the strings of numbers.

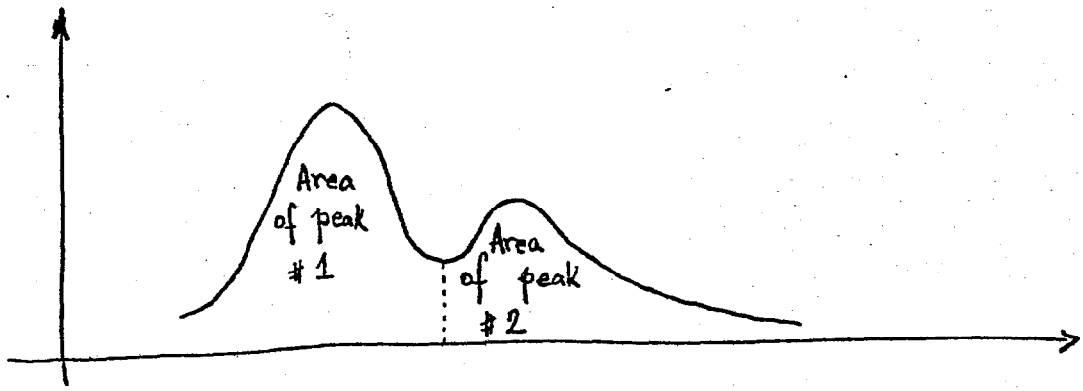
(3) Once a peak is coded in terms of the coefficients of its associated approximating polynomial, this code is compared with a "library of peak shapes" and a call is issued to an appropriate peak resolving algorithm, specifically designed to deal with this type of shape.

(4) In this last step the referenced algorithm finally performs the task of peak resolution, using as an input the peak parameters determined in the step number 2. Having in mind that the performance of most of the peak resolving algorithms depends crucially upon the correct initial guess of the peak parameters, we may surmise that a good overall performance of the algorithm will depend a great deal upon the cleverness of the "preparation" steps 1, 2 and 3. We recall that the proper functioning of the peak resolving algorithms of the least squares type is particularly sensitive upon a good first guess of the peak parameters.

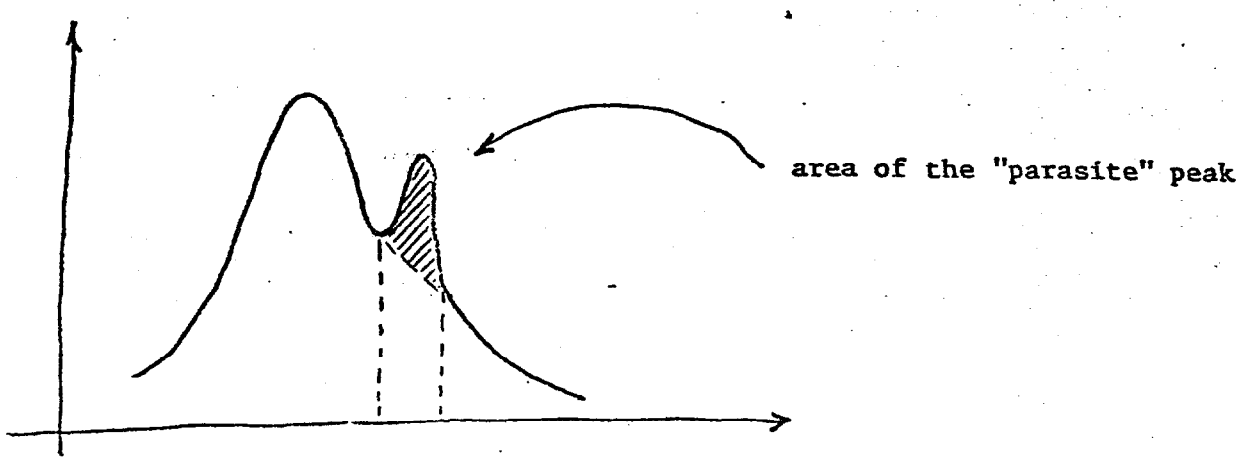
I was unable to find in the literature a reference to an algorithm of the described type. Instead, one finds very simple algorithms, intended to deal with equally simple cases of composite peak shapes. Few examples are given below.

Geometrical Method:

If two peaks are sufficiently removed in their overlap to generate a dip in the middle, some simple approximate geometrical methods for peak resolution can produce satisfactory results. For example, one can drop a straight line from the lowest point in the dip and calculate the individual areas as illustrated by

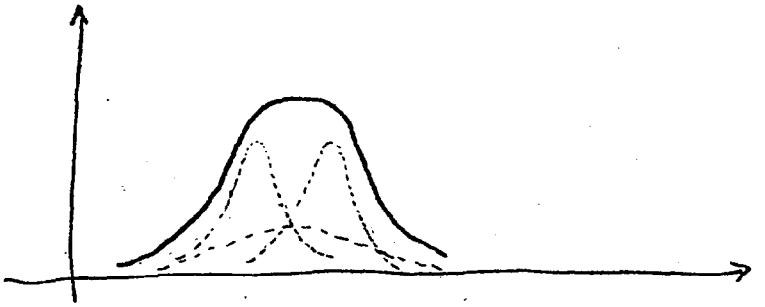


or, if a small peak "rides" on the tail of a larger peak, one can employ so-called "tangent skim" method depicted below



TANGENT SKIM METHOD

Evidently this method is not applicable in the situations where the local dip is missing, as the following "difficult" case nicely exemplifies:



One special case:

In certain circumstances the locations and the shapes of the elementary peaks are known, while their relative intensities in the composite peak are not. The problem is, then, to determine the coefficients c_i in

$$f(x) = c_1 P_1(x - x_1) + c_2 P_2(x - x_2) + \dots + c_n P_n(x - x_n)$$

where $P_i(x - x_i)$ are the functions of the elementary peaks and X_i their locations.

The method of least squares requires that a set of c_i's be found which forces the integral

$$I(c_1, c_2, \ldots, c_n) = \int \left[f(x) - c_1 P_1(x-x_1) - c_2 P_2(x-x_2) - \cdots - c_n P_n(x-x_n) \right]^2 dx$$

to reach its minimum. The mathematical expression of this statement is given by

$$\frac{\partial I}{\partial c_i} = 0 \qquad , \; i = 1, 2, 3, \ldots, n \; ,$$

a set of conditions which leads to the following system of n linear equations with n unknowns

$$\int f(x) P_i(x-x_i)\,dx = c_1 \int P_1(x-x_1) P_i(x-x_i)\,dx + c_2 \int P_2(x-x_2) P_i(x-x_i)\,dx + \cdots$$

$$\cdots + c_n \int P_n(x-x_n) P_i(x-x_i)\,dx \; ,$$

$$i = 1, 2, 3, \ldots, n \; .$$

The solution of this system gives the answer to the problem of the composite peak resolution for this special case of known locations and shapes of the composite peak components.

Convolution:

$f(x)$ denotes, as usual, the function that describes the shape of a composite peak. One constructs the integral

$$\int f(x)\rho(x-y)\,dx \quad , \qquad \int \rho(z)\,dz = 1$$

in the hope that with the judicious choice of the known test function $\rho(z)$, a useful information can emerge when the variable y scans the region of the peak. In the extremal situation when the test function is the δ – function, the integral reduces to the identity

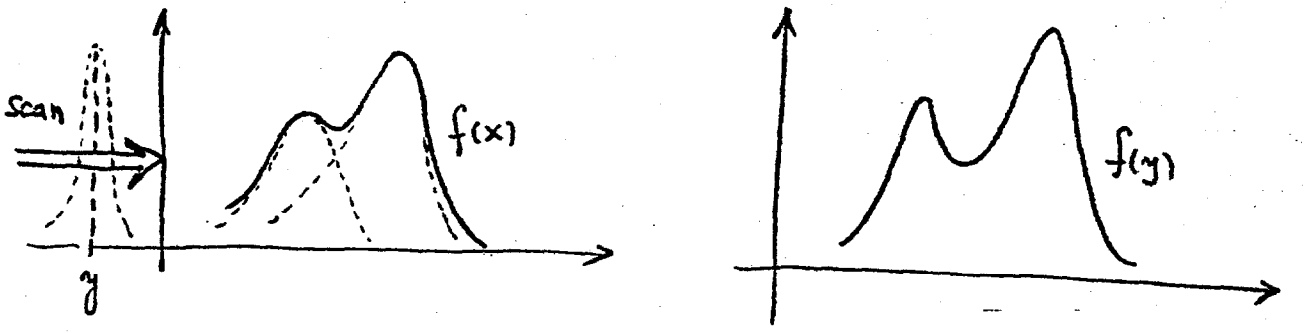
$$\int f(x)\,\delta(x-y)\,dx = f(y) \quad ,$$

and no new information can be extracted from the convolution. On the other end of the spectrum of possibilities ρ is a constant, and we have

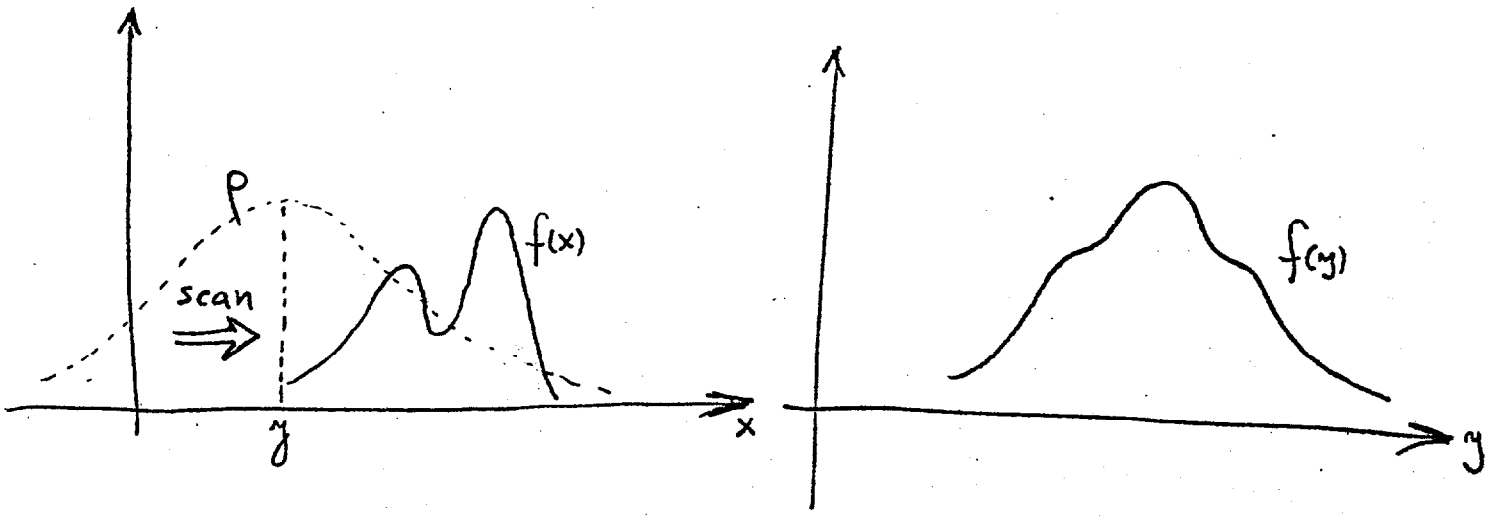
$$\int f(x)\rho(x-y)\,dx = \text{const.}$$

Again, very meager information is obtained.

For the success of the convolution method it is absolutely essential that a correct choice of ρ is being made. The best results will be obtained when a composite peak consists of a sequence of elementary peaks of approximately identical widths, and the test function matches this width with its own width. Then, whenever the test function passes over the one of the elementary components of $f(x)$, an appropriate enhancement occurs with better resolution as a result. This situation is depicted below.

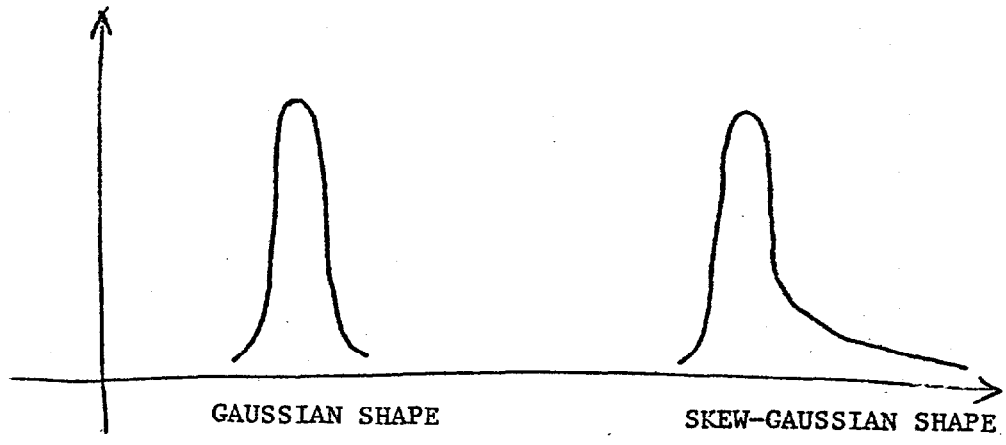


When using the convolution method one should stay away from the choice of too broad test functions, which may result in a misleading information. This is illustrated by the following example:



Are the peaks Gaussian?

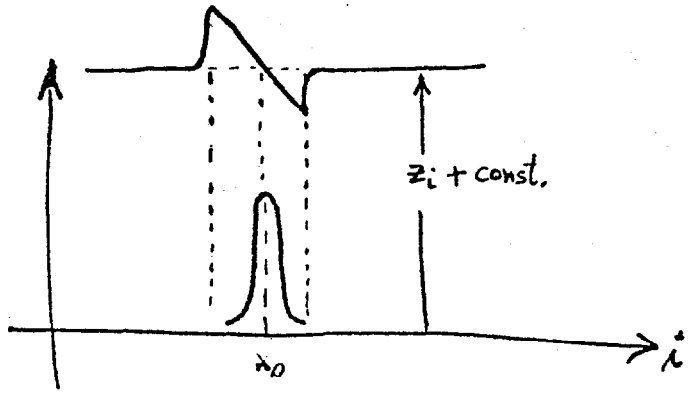
There seems to be no acceptable theory, based on first principles, which predicts the shape of peaks observed in the gas chromatography. In the literature one finds claims [4] that in the ideal experimental circumstances the peaks should be gaussian, but in practice this is rarely, if ever, realized. At best, the peaks display a regular shew-gaussian form, for which one can find an appropriate analytical expression [5].



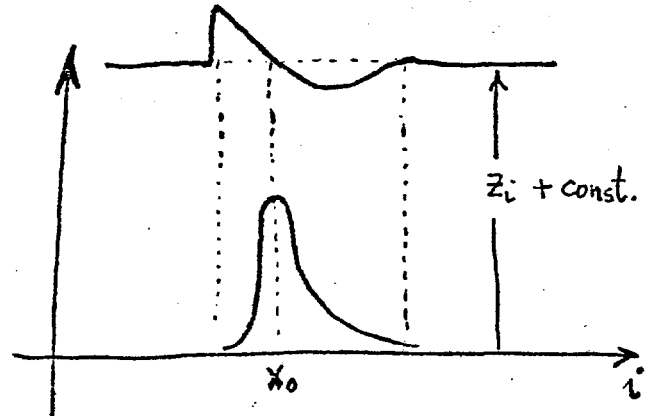
There exist a very simple test for checking whether a peak is gaussian. It is based on the characteristic property of gaussian functions, namely

$$\left[e^{-b(x-x_0)^2} \right]^{-1} \frac{d}{dx} \left[e^{-b(x-x_0)^2} \right] = 2b(x_0 - x)$$

The associated algorithm goes as follows. Denote by y_i the data amplitudes. Construct the array $z_i = \dfrac{y_{i+1} - y_i}{y_i}$ and plot z_i versus i. Over a genuinely gaussian peak this plot should be linear.



GENUINE GAUSSIAN PEAK



ACTUALLY OBSERVED PEAKS

Unfortunately, it does not appear that this simple method can be modified so that it can be used in the problem of resolution of composite peaks. The deviations from linearity tell us very little about the internal structure of the composite peaks - except that they are definitely not gaussian.

Other types of peak shapes:

Aside from gaussian functions, other analytical forms have been tried, some of them quite involved [5] and depending on as much as four or five free parameters. We do not list here all options that are offered in the literature, but only briefly discuss the Poisson Distribution Peaks [6] which seem to be of relevance in gas chromatography. The functional form is determined by the two parameters a and n :

$$P_{n,a}(x) = c(n,a)\, x^n e^{-ax} \quad , \qquad c \text{ is a normalization constant.}$$

It is easy to verify that the function $P_{n,a}(x)$ satisfies the differential identity

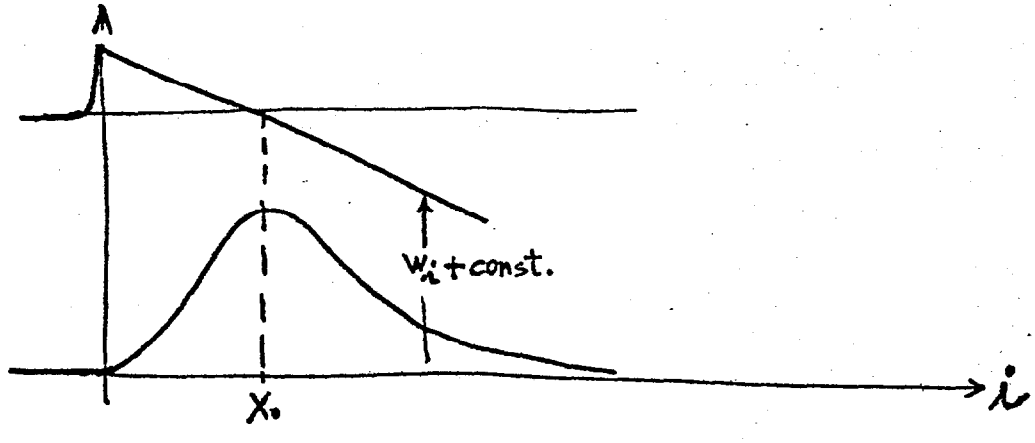
$$x\, P_{n,a}^{-1}(x) \frac{d}{dx} P_{n,a}(x) = n - ax \quad ,$$

and it is this identity which serves as a basis for the following simple test (whether a peak is pure $P_{n,a}(x)$ type or not).

Let, again, y_i denote the data amplitudes. Construct the array

$$w_i = i \frac{y_{i+1} - y_i}{y_i}$$

and plot its points versus the variable i. If the peak under inspection is of pure Poisson Distribution type, the array points w_i over the peak will lie on a straight line, as illustrated by



PURE POISSON DISTRIBUTION PEAK

The idea for this test is obviously a slight modification of the idea for testing a gaussian character of a peak, and shares with the latter all virtues and faults. In particular, it is difficult to imagine what useful information one could extract from a discovery that w_i's do not lie on a straight line - except that in this case the peak is definitely not of Poisson Distribution type.

The peaks of Breit-Wigner shape may also be of some interest in gas chromatography. The shape, location and the peak size are determined by three parameters, a, b and x_0:

$$BW_{a,b}(x) = \frac{a}{(x - x_0)^2 + b^2}$$

If a peak is of a pure Breit-Wigner type, the points of the array

$$t_i = \frac{y_{i+1} - y_i}{y_i^2}$$

should lie on a straight line, when the subscript i runs over the area of the peak.

SUMMARY

A large number of various approaches to the problem of peak resolution are known today. It is safe to say that none of these approaches can claim a complete success when confronted with composite peaks of sufficient complexity (for example, an overlap of four or more peaks). The best strategy, then, seems to be to avoid the occurance of multiple peak clusters altogether. The novel approach to the utilization of data from the GC-MS combination, mentioned in the Introduction, offers a great promise in this respect. Since it is quite unlikely that more than two chemical substances in a sample, with very similar GC retention times, will have only one prominent ion in their mass spectra, and that this ion will happen to be the same, we can expect to encounter rarely, if ever, the need to deal with an overlap of more than two peaks in an individual mass chromatogram. This situation should be contrasted to a total ion current plot from a gas chromatograph, where an overlap of three or more peaks is not an infrequent occurance. But an overlap of two, or even three, peaks is sufficiently simple to be amenable to analysis by most good peak-resolving algorithms. My preference, in this case, is either the use of standard least squares techniques [5, 7], or a good heuristic algorithm based on geometrical considerations .

REFERENCES

1. J. E. Biller and K. Biemann, Analytical Letters 7 (7), 515 (1974).
2. T. Rindfleisch, private communication.
3. G. Dromey, in preparation
4. J. C. Giddings, "Dynamics of Chromatography, Part I: Principles and Theory", Marcel Dekker, Inc., New York, e.Y., 1965.
5. H. M. Gladney, B. F. Bowden, and J. D. Swalen, Analytical Chemistry 41, 883 (1969).
6. E. Grushka, M. Myers, and J. C. Giddings, Analytical Chemistry 42, 21 (1970).
7. See, for example, J. T. Routti and S. J. Prussin, Nucl. Instrum and Methods 72, 175 (1969).

Distribution:

A. Duffield

G. Dromey

W. Pereira

T. Rindfleisch

D. Smith

M. Stefik